DOCUMENT RESUME

ED 173 445 TM 009 623

AUTHOR Bergquist, Constance C.
TITLE The Reliability of Auditor Judgments: A Meta-evaluation Study.
PUB DATE Apr 79
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Decision Making; Evaluation Criteria; *Evaluation Methods; *Evaluators; Program Administration; *Program Evaluation; *Reliability
IDENTIFIERS Auditors; *Audits; Florida; *Interrater Reliability

ABSTRACT

The reliability of evaluators' judgments was investigated, using a discrepancy evaluation model within a program evaluation context. Two research problems were addressed: the degree of evaluator reliability in detecting and rating problems in federal educational project planning, management, and evaluation; and the relationship between inter-judge reliability and the types of items examined, the evaluators' expertise, and the type of project being evaluated. Of thirty-one evaluators studied, 77% had doctoral degrees, and all had previous training related to the Title IV-C audit system. Three types of reliability coefficients were calculated: (1) the consistency of a single rater over a set of ratings was .07 to .64; (2) the estimated single rater reliability over all situational protocols was .20 to .84, with most coefficients above .50; and (3) the average interrater reliability was .45 to .96, with over half of the coefficients above .80. Results raised two questions about the reliability of judgments made by educational evaluators; questionable replicability of evaluation methods and standards, and the amount of disagreement among evaluators on methods and procedures for evaluating programs. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED173445

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Reliability of Auditor Judgments: a Meta-evaluation Study

by

Constance C. Bergquist, Ph.D.
Evaluation Systems Design, Inc.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Constance C. Bergquist

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979

TM009 623

The Reliability of Auditor Judgments: a Meta-evaluation Study

The reliability of the decisions or judgments of evaluators is essentially an uninvestigated phenomenon. Although the profession has consistently recommended strict application of empirical research principles to the conduct of evaluation activities, little effort has been made to apply the same principles to the profession. In evaluation research there are few empirical studies related to the reliability of evaluator judgments, even though funding of projects amounting to thousands or millions of dollars depend on the professional judgments of evaluators. The purpose of this study was to investigate the reliability of judgments made by evaluators using a discrepancy evaluation model within a program audit context.

Statement of the Problem

As part of an extensive planning and management system to ensure the development and implementation of exemplary Title IV-C innovative projects, the Florida Department of Education, Educational Innovations Section, developed a Standardized Educational Program Audit (SEPA) that was implemented in 1974. Although a formative examination of the auditing system was conducted in 1974, no studies had been conducted to determine the reliability of the auditors in detecting and rating project problems.

Two specific research questions were addressed in the study. First, how reliable are evaluators as auditors in detecting and rating problems in federal educational project planning, management, and evaluation? This question addressed the need for empirical evidence concerning evaluator/auditor reliability. Data on auditor reliability were needed by the Florida Department of Education, Educational Innovations Section, to facilitate evaluation of the auditing system for Title IV-C projects.

Second, does inter-judge reliability depend on the types of audit items examined, the expertise of the auditor, the type of project being audited, etc.? These questions related to the identification and, where possible, the investigation of factors that may affect the reliability of auditor judgments.

Method

The stimulus for this study was a series of simulations of audit settings to which auditors responded independently. Simulations were developed for two projects, a second-year developmental project and a second-year demonstration project. The simulations contained descriptions of the project purpose, objectives, progress-to-date, personnel, facilities, and clients. Instruments utilized to collect the data were the actual audit protocols for the corresponding item stimulus. Responses elicited were the classification of the audit findings and the rating of the risk/recovery factors for any identified inconsistencies or discrepancies.

The subjects of the study were 31 auditors with previous training in the Title IV-C audit system. The educational level of the group was extremely high with 77% of the auditors having received doctoral degrees. More than half of the group were professors while a large percentage worked in local school districts. Areas of specialization included research design, measurement, statistics, curriculum development, and instructional design.

For the data analysis, individual rater and average rater reliabilities of the ratings by the auditors were estimated through the application of analysis of variance. This procedure has been described and discussed by Ebel (1951), Guilford (1954), and Winer (1971). Since a sample of four protocols was taken for each type of auditor and project, the individual reliabilities were underestimates of the reliability of the judgment of a single auditor over all protocols within his/her auditing type. Therefore, an extrapolated individual rater reliability estimate was calculated. Individual protocol data were also summarized and analyzed to determine the protocols on which the least consistency was found.

Results

Three types of reliability coefficients were calculated for the study. A sample of the coefficients calculated is presented in Table 1. Table 1 displays the reliability coefficients calculated for management and program auditor judgments on both developmental and demonstration projects. Other coefficients calculated were for various combinations of auditor type and project type.

Table 1

Reliability Coefficients for Management and Program Auditor Judgments

Reliability Coefficients	Management Auditors (N=16)		Program Auditors (N=15)	
	Development Project	Demonstration Project	Development Project	Demonstration Project
Individual reliability coefficient	.28	.24	.58	.64
Estimated reliability coefficient for entire protocol set	.60	.46	.82	.84
Average reliability coefficient	.87	.84	.96	.96

The first coefficient calculated, designated the individual rater reliability coefficient, provided an estimate of the reliability or consistency that could be expected for the set of ratings completed by a single rater. The range of these coefficients was from .07 to .64. By any standards these coefficients were very low.

The second type of reliability coefficient calculated was an extrapolation of the first coefficient to reflect the estimated reliability of the judgments provided by a single rater over all of the protocols, both sampled and unsampled. The ranges of the coefficients were from .20 to .84, with most of the coefficients above .50.

The third type of reliability coefficient calculated was the average inter-rater reliability coefficient, which provided an estimate of the reliability of the mean rating calculated from all of the auditors responding to the simulation. This coefficient is relevant only if decisions were to be made on the basis of the mean rating of several auditors. The range of the average rater coefficients in this study were from .45 to .96 with more than half of the coefficients above .80. Generally, these coefficients were quite high by the standards typically established in the field of educational research.

Comparison of the reliability coefficients calculated for program auditor judgments with management auditor judgments indicated that judgments made on program protocols were more reliable than judgments made on management protocols. These differences may reflect the degree of complexity in the management and evaluation protocols. They may also reflect the diverse opinions in the field of evaluation on the appropriateness of some research design and statistical analysis procedures.

Examination of the effect of experience on the auditor judgments indicated that experience increased the reliability of the management auditor judgments but did not increase the reliability of the judgments of program auditors. Examination of the effect of the type of project on the reliability of the judgments revealed that coefficients for management auditors were higher on development projects than on demonstration projects, but program auditor judgments were more reliable on demonstration projects than development projects.

The individual protocol data were also analyzed. Means, standard deviations, frequency distribution of responses, and percentage of responses by scale score were calculated for each protocol included in the simulation. The criterion adopted to determine the consensual level of agreement on the individual protocol responses was that at least 70% of the responses had to fall within 2 of the 5 scale points in order for the protocol responses to be considered reliable.

Tables 2 and 3 present for each protocol the percentage of responses by scale score, the highest cumulative percentage falling within two scale categories, and results of application of the criterion to the cumulative percentage of responses. Only four of the eight management protocols met the established criterion, but seven of the eight program protocols met the criterion. The criterion of 70% consensus was probably the minimum level that would be acceptable in the field of educational evaluation.

Table 2

Percentage of Responses, Criterion Levels, and Criterion Decisions for Management Protocols

	Protocol		Response Category					Criterion Frequency	Criterion Decision
			NA	Consis-tent	Minor	Prob-lem	Crisis		
			0	1	2	3	4		
1.	1.1.1	Work Plan Implementation(N=15)		53		20	27	53	
2.	3.1.2	Structure of Design (N=13)		8	15	15	62	77	yes
3.	3.2.3	Instrumentation (N=10)		10	10	50	30	80	yes
4.	3.3.3	Data Analysis (N=10)		10	30	30	30	60	
5.	1.1.2	Monitoring/Sign. Events (N=15)		13	20	33	33	67	
6.	1.2.1	Communication (N=16)		6		50	44	94	yes
7.	3.2.3	Program Usage (N=16)		13	38	38	13	75	yes
8.	3.2.4	Usage Results (N=16)	31	6	13	19	31	50	

Table 3

Percentage of Responses, Criterion Levels, and Criterion Decisions for Program Protocols

	Protocol		Response Category					Criterion Frequency	Criterion Decision
			NA	Consistent	Minor	Problem	Crisis		
			0	1	2	3	4		
1.	1.1.4	Monitoring/Prod.Status (N=14)			7	71	21	92	yes
2.	2.1.1	Delivery System Design (N=14)		64	21	7	7	86	yes
3.	2.2.2	Technical Review (N=14)		7		36	57	93	yes
4.	2.2.4	Small Group/Pilot Test.(N=14)				50	50	100	yes
5.	2.1.3	Inventories (N=14)		14	57	21	7	78	yes
6.	2.2.1	Consumer Awareness (N=15)		20	40	27	13	67	
7.	2.2.3	Consumer Training (N=14)		50	36	14		86	yes
8.	3.1.2	Response Time (N=14)				50	50	100	yes

Discussion and Conclusions

The results of the study indicated that estimated coefficients for audits of the entire set of protocols by individual auditors ranged from .20 to .84 with most of the coefficients above .50. These low coefficients raised two serious questions concerning the reliability of judgments made by educational evaluators. First, taken in the context of program evaluation, these results indicate questionable replicability of evaluation methods and standards. If 40% of the auditors, as experts in evaluation, consider the use of analysis of covariance appropriate even when random sampling is not employed, and 60% of the auditors rate this data analysis as a problem or crisis situation, the credibility of the profession may be seriously affected.

The second question raised was in the use of "expert" review to establish the validity of instruments or procedures in the area of educational evaluation. If the experts cannot agree on methods and procedures for evaluating programs, as indicated by the low reliability coefficients, the benefit of expert review should be questioned. The low reliabilities cast doubt on the validity of the use of individual expert judgments for program review.

Although these estimates of the reliability of judgments made by individual auditors were low, they may be high enough for use in a comprehensive grant management system. In the Florida Title IV-C system, the audit is only one method of reviewing project progress. Results from the audit are subjected to an extensive review and appeal process that

may reduce the level of reliability necessary in some areas of program management. If the audit were used as the final determination of project worth or accomplishment, higher levels of reliability would probably be necessary.

The coefficients calculated reflecting the reliability of the mean rating from all auditors ranged from .45 to .96, with more than half of the coefficients above .80. These coefficients indicate the greater reliability of systems which employ several auditors independently judging program success or progress. Using the mean of these judgments results in higher reliability. An example of this kind of program review would be the Joint Dissemination Review Panel. Although using the mean of the judgments of several auditors may increase the reliability, it also greatly increases the cost of the system and may not be worth the additional expense.

Evaluations of the simulations used in determining the reliabilities indicated that the reactions of the evaluators were quite positive. These evaluations indicated the potential utility of the technique as a training setting for novice auditors.